Opening the Black Box of Play: Strategy Analysis of an Educational Game

Britton Horn
Northeastern University
Boston, Massachusetts
bhorn@ccs.neu.edu

Amy K. Hoover
Northeastern University
Boston, Massachusetts
amy.hoover@gmail.com

Jackie Barnes
Northeastern University
Boston, Massachusetts
jacqbarn@gmail.com

Yetunde Folajimi
Northeastern University
Boston, Massachusetts
yetundeofolajimi@gmail.com

Gillian Smith
Northeastern University
Boston, Massachusetts
gi.smith@northeastern.edu

Casper Harteveld
Northeastern University
Boston, Massachusetts
c.harteveld@northeastern.edu

ABSTRACT

A significant issue in research on educational games lies in evaluating their educational impact. Although game analytics is often leveraged in the game industry, it can also provide insight into player actions, strategy development, and the learning process in educational games separate from external evaluation measures. This paper explores the potential of game analytics for learning by analyzing player strategies of an educational game that is designed to support algorithmic thinking. We analyze player strategies from nine cases in our data, combining quantitative and qualitative game analysis techniques: hierarchical player clustering, game progression visualizations, playtraces, and think-aloud data. Results suggest that this combination of data analysis techniques provides insights into level progression and learning strategies that may have been otherwise overlooked.

ACM Classification Keywords

H.5.1 Information Interfaces and Presentation (e.g., HCI): Multimedia Information Systems—*Evaluation/methodology*; K.3.2 Computers and Education: Computer and Information Science Education—*Computer science education*; K.8.0 Personal Computing: General—*Games*

Author Keywords

educational games; computational thinking; playtrace analysis; evaluation.

INTRODUCTION

Through the emerging field of *game analytics* [8] (analogous to *learning analytics* [37]), methods for both evaluating gameplay and analyzing player behavior have been dramatically expanding [27]. Rather than evaluating learning outcomes through external evaluations alone (e.g. , questionnaires and tests), recent approaches investigate player actions logged during gameplay to provide insights into the learning processes of players [25, 12, 13]. A significant challenge in game analytics and educational game design lies in successfully combining traditional methods of assessing knowledge with tracking and analyzing behavioral telemetry for the purposes of both assessing learning and improving the design of educational games.

This paper contributes to analyzing educational games with game analytics by exploring player strategies in *GrACE*, an educational puzzle-based game for middle school aged students (11-13 years old) that is designed to support algorithmic thinking [20]. While many educational games such as *CodeCombat* [4], *Robocode* [29], and *Robozzle* [24] also focus on supporting algorithmic thinking, they often emphasize the structure and syntax of computer programming rather than addressing the core planning and strategy development processes. *GrACE* on the other hand, supports algorithmic thinking by presenting players with a puzzle analogous to solving a typical computer science problem (i.e., finding the minimum spanning tree). Through navigating and solving multiple puzzles, the goal is for players to learn both the data structure and the step-wise algorithm for solving it.

One obstacle in game analytics is that it is often difficult to correctly interpret the meaning of player data [18]. To interpret the meaning of player actions in *GrACE*, we take the approach of triangulating findings by combining quantitative hierarchical cluster analysis of player actions with a qualitative analysis of playtraces supported by concurrent think-aloud data and progression visualizations. We found that our mixed-methods approach helped discover emergent player strategies and how the game mechanics may have supported such strategies, findings that may not have been visible through traditional assessments or game analytics alone. Our work unveils prevalent player strategies by "opening the black box of play" using this triangulation approach, building a deeper understanding of how players learn and progress in the game, and aiding

CHI PLAY '16, October 16-19, 2016, Austin, TX, USA

 ISBN 978-1-4503-4456-2/16/10…$15.00

DOI: http://dx.doi.org/10.1145/2967934.2968109

decisions in the (re)design process of building an effective educational game.

BACKGROUND

This section elaborates on why game analytics should be considered when developing and evaluating educational games. It also discusses existing playtrace analysis techniques used in commercial and educational games.

Game Analytics for Educational Games

Although there is a long and rich history of developing games for education, such games often lack rigorous evaluation [14]. While used in game development, evaluating learning using game analytics is relatively new. Given time constraints of educators or researchers to distribute and evaluate traditional assessments, or code interview data, game analytics is a promising alternative to efficiently and automatically evaluate performance and learning. In fact, some scholars argue that game data itself is an assessment of learning (provided that appropriate metrics have been defined) [36]. Shaffer and Gee [35] even state that "We have been designing games for learning when we should have been designing games for testing" (p. 3). As evidence of how game data can act as an assessment of learning, one recent study showed a strong correlation between game performance and external test measures [17].

Further, analyzing game data should not only determine whether a design works, but also *how* it works. For example, in a study that analyzed both efficacy measures as well as game data [15], it was found that the game was an effective training intervention, but looking closely at the playtraces indicated that players consistently made the same errors, suggesting the design itself could be improved. Asking only efficacy questions can discount other aspects of player experience and potentially result in incorrect conclusions. Given this, game analytics can facilitate data-driven game development, which can expose design problems at an early stage. Although such data-driven design is becoming increasingly common in the industry [8], it is not yet a common practice for educational games. Data-driven game development has the potential to assist in the formative evaluation of achieving learning objectives [14].

Tools and Methods for Playtrace Analysis

Many methods have been used for playtrace analysis [10], including traditional observational studies [19] as well as videotaping play and interview sessions [1], all of which are qualitative in nature and are difficult to use in analyzing large-scale playtest data [3]. In contrast, statistical and machine learning techniques have been used to track and categorize players in, among others, *Bioware* [6], *Forza Motorsport* [33], and *World of Warcraft* [7] by leveraging raw game data. For these quantitative approaches, scholars developed and validated aggregate metrics from the raw data to measure game states or player behavior [8], such as the number of attempts or level difficulty.

Using such metrics to describe player behavior has also been applied to educational games. For example, Serrano-Lugano et al. [34] introduce a scalable two-step approach where they first define simple generic metrics that could be applied to any educational game and then build game-specific assessment rules based on combinations of these metrics. They found that their approach is valid for identifying where players get confused, but it does not give insight into *how* to address player confusion, likely due to a lack of understanding of why players behave in a certain way [18].

In addition, to better understand player decisions during gameplay and how player choices differed from what designers expected, a playtrace of an educational game has been done using conceptual feature extraction methods drawn from log data [12]. Researchers created various models that break gameplay down into individual cognitive tasks (or knowledge components) and assess which of these models is the best descriptor of player learning throughout the game. Though this work aims to gain insights into the dynamics of learning in an educational game, it relies solely upon logged game data and performing statistical regressions to better understand predictor variables.

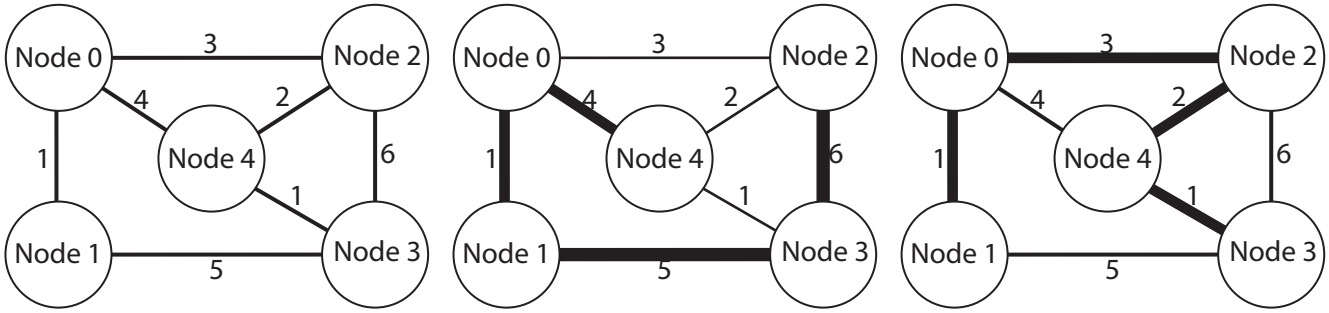
To decode *how* players play games, tools have been developed to visualize player trajectories. In fact, two existing efforts, *Playtracer* [3, 26] and *Glyph* [28], have both supported the qualitative and quantitative analysis of an educational puzzle game through state graphs to identify common play patterns. *Playtracer* calculates the distance of states to the goal, whereas *Glyph* uses defined player actions as the edges connecting the game states. Even with these tools, it is a challenge to decide how to define game states, their relationships to one another, and to interpret the data being visualized.

Historically, playtrace analyses have developed from small-scale qualitative studies to large-scale quantitative studies. The latter was enabled with the arrival of game analytics and the need to study larger data sets; however, this came at the cost of meaning, which has been the strength of traditional playtrace studies. In our work, we pursued a mixed-methods approach to address the issue of interpretation in analyzing player strategies in *GrACE*.

DESIGN

GrACE is a puzzle game with a vegetable-collecting narrative designed to encourage algorithmic thinking. Algorithmic thinking involves thinking about problems abstractly, identifying common traits so that they can be treated as a class of problems instead of a single instance, and building sequences of instructions as solutions. It is a logical method of problem solving that can also be applied in areas outside computing [40]. Because *GrACE* is being developed in part to interest girls in computer science, the puzzle genre is chosen based on research suggesting it is appealing to girls [11, 31]. Furthermore, the puzzle genre is one of the top two genres played by teens [22] and like much of the computer science discipline, is about logically solving complex problems [16].

Specifically, *GrACE* aims to illustrate the potential of teaching algorithmic thinking through puzzles analogous to finding the minimum spanning tree (MST) of a graph, a core problem in computer science. These MST-based puzzles were chosen over other possibilities for their ease of visual representation, the existence of many algorithms to solve the problem, and



(a) Graph example (b) Spanning tree example (c) Minimum spanning tree example

Figure 1. Each connected graph such as the example in (a) has a variety of spanning trees, including the ones in (b) and (c). Because the cost of the path illustrated in (b) is 16 and that shown in (c) costs 7, (c) represents a less costly path and in this case represents the minimum spanning tree.

that even finding an incorrect solution (i.e., a spanning tree that is not minimal) involves algorithmic thinking.

Abstractly, the MST-based puzzles in *GrACE* are represented as graphs, which are a collection of nodes connected by edges. Each edge has an associated cost as shown in Figure 1(a), which shows an abstract MST puzzle. Players travel from one node to another along the edges that connect them. The cost of including an edge in the solution is represented by the number above the traveled edge. For example, in Figure 1(a), the cost of the edge between Node 0 and Node 2 is three.

In these graphs, players can travel along any edge connected to the node they are currently visiting. To indicate a particular edge should be considered in the solution of the puzzle, players can flag the edge when visiting the nodes connected to it. The darkened edges in Figure 1(b) like the edge connecting Node 0 and Node 4 represent such flagged edges.

A "spanning tree" is formed when a player visits all of the nodes and flags edges such that each node is only connected to the others once (as shown in Figure 1(b)). While a graph can have many different spanning trees, those with minimal cost are a MST (shown in Figure 1(c)). The abstract goal in *GrACE* is for players to find a MST by traveling along and flagging the least costly path where each node is connected to the others exactly once.

To avoid computer science jargon and create an interesting and relevant context, solving the MST-based puzzles centers around a narrative with two characters, Scout the mouse and Hopper the rabbit, whose combined goal is to collect all possible vegetables while expending the least amount of energy. In this narrative, burrows represent the graph nodes where each burrow contains a vegetable. The graph edges are tunnels that Hopper can dig to collect the vegetables. The cost of each tunnel is represented by the number of rocks through which Hopper must dig.

In the game, players control the character Scout, who traverses the graph while simultaneously placing flags on the edges to indicate which tunnels the character Hopper should dig. Scout the mouse is too small to dig but is able to traverse the graph unlike Hopper the rabbit. Scout shows Hopper where tunnels

Figure 2. Screenshot of *GrACE* ©Northeastern University. This puzzle currently shows three nodes (as other nodes may be revealed later) with two edges, both with a cost of one, indicated by the arrow and number of rocks. Flagging involves literally placing a flag in the middle of an edge. On the top right are Hopper's and Scout's energy levels, to the left are the level select options, and at the bottom the submit button.

should be dug by flagging paths. At any point during the game, players may place or remove a flag on an edge.

Once players think they have explored the graph sufficiently and found the best possible solution, Hopper will dig tunnels along the flagged path, and collect the vegetables. The total cost of the flagged paths represents the amount of work for Hopper to dig tunnels, and if this is more than Hopper's available energy (represented as an energy bar) then players have not found an MST, which is necessary to successfully complete the puzzle and proceed to the next level.

Scout also has an energy bar which decreases with every action the player takes (moving, flagging, unflagging). For completing the level, Scout's amount of energy is not relevant except if she runs out, because in that case players will have to try again. Scout's energy has been included to encourage players to explore puzzles efficiently, and thereby algorithmically.

To simulate the way computers "think" through the problem node by node while keeping track of discovered edges and their weights, players initially can only see the starting node and the nodes which are directly connected to it. Players choose an edge to traverse at which point the new node and its

connections are revealed. By slowly revealing the entirety of the graph, players are encouraged to develop an algorithmic approach which helps them solve more complex problems that would be challenging with only visual inspection.

Once all nodes have been explored and desired edges flagged, the player may submit their answer by clicking the "Submit" button. A correct answer sends the player to the next level which will contain more nodes and edges than the previous, while an incorrect answer leaves several options: the player may edit their submitted answer (which will continue to reduce Scout's available energy), restart the level, or replay any previously successfully submitted puzzles.

METHOD

To both explore how well *GrACE* encourages algorithmic thinking and to evaluate the ability of players to solve MSTs, this paper analyzes game data collected from a pilot study with *GrACE* [20]. Results from traditional external evaluation measures indicated a moderate improvement in scores after playing. Gameplay metrics were also considered (e.g., time for completion of each level, the number of correct and incorrect submitted solutions, and the number of levels completed), and indicated that participants tended to stumble at particular levels, with some never completing a level past the fifth (out of 11). Because some participants completed all eleven puzzles and some did not, a difference in strategies (and learning) may explain the variety in progress. To analyze players' strategies, we selected a small sample of participants and performed an in-depth analysis using qualitative and quantitative techniques.

Our approach combines a quantitative hierarchical clustering of player actions over the course of the game with a qualitative analysis of playdata that we call *retrospective player sense-making* collected through concurrent think-aloud data. Retrospective player sense-making involves reconstructing what a player did in the game by examining their playtrace step-by-step, noting observations along the way. Cluster analysis in particular is susceptible to issues of interpretation and trial-and-error [38, 39], so the advantage of this combined approach is that findings can be triangulated. Also, the think-aloud data may provide insights into players' in-the-moment perceptions and goals. Playtraces alone only help infer *what* players did rather than *why* they did it. Overall, our goal was to understand the divergence of player trajectories beyond quantitative game data. That is, we wanted to get an idea of why some players got stuck on certain levels and others were able to progress easily.

Study Context

The pilot study [20] was implemented as a programmed three-hour activity in a two-week summer program at Northeastern University that focuses on STEM education and retention for underserved and underrepresented minorities. Talented middle school students (ages 11 through 13) are selected to participate in this free of charge program. The original study was a 2x2 experimental design, where students received a prequestionnaire and pretest at the start. Following an explanation of how to play the game (without mentioning MSTs or the game's relationship to computer science), participants were randomly assigned to one of the conditions and played the game for about an hour. If students finished playing all levels, they were encouraged to repeat past levels and attempt to minimize the energy used by Scout the mouse. The activity ended with a postquestionnaire and posttest, followed by a group discussion.

Selection of Case Studies

Out of the 43 students who participated with the consent of their parents in the pilot study, we selected nine. These nine students participated in the same condition and played the same puzzles making it possible for us to perform an in-depth comparative analysis on how they played. For these selected students, ages ranged from 11 to 13 (M = 12.2, SD = 1.06). Four students identified themselves as female and five as male. Additionally, one student identified as Hispanic or Latino, three as Asian, one as Black or African American, two as White, and one as "other" (with one participant preferring not to answer). Each student indicated interest and experience playing games (e.g. video games, board games, sports), with some indicating that they played games several times a day while others only played every few weeks. Therefore, the selected nine students were diverse in terms of their socio-demographics and frequency of play (see also Table 1).

Data Collection

Of relevance to this paper is that players were asked to complete a postquestionnaire on their game experience with *GrACE* as well as a pretest and posttest to assess their algorithmic thinking. The postquestionnaire measured play experience with four 7-point Likert items focused on enjoyment and four on difficulty. The resulting composite scores range from 4 to 28, and are both reliable and valid [20].

The tests were developed to measure conceptual understanding of the MST and involved a multiple choice assessment of nine test questions with four choice options. All test questions were phrased in the context of a specific MST puzzle such as deciding what edge to add, what edge to remove, and what needs to be corrected. Test scores were calculated by providing one point per correctly answered question. Thus, with nine questions students could get a maximum score of nine points. The pre- and posttest differed in the exact questions asked, but were the same type.

The game data logged for each player was collected and each player also received a USB voice recorder to record their talk while playing. We encouraged students to think-aloud while playing so we could infer from their talk why they are playing in a certain way.

Analysis

The analysis considers players' strategies and progression through cluster analysis and retrospective player sense-making. For the cluster analysis, each playtrace was first converted to a string representing the sequence of actions performed. For instance, if a player successfully completed a level by starting at Node 0, moving to Node 1, flagging that edge, and then submitting their answer, these actions would be represented as "Start Node0, Move Node0 Node1 Edge1, Flag Edge1, Submit Correct." These strings are then compared by calculating the

Levenshtein string edit distance between them [23, 30], which counts the number of additions, substitutions, and deletions necessary to convert one string to the other. This number indicates the similarity between strings of playtraces.

For the retrospective player sense-making, we considered player strategies for each level and how these strategies evolved as players attempted each new level. We also used level progression visualizations to observe where players got stuck and puzzle features that may explain why. Once patterns emerged, we worked to better understand these patterns by supplementing initial findings with in-the-moment audio recordings of players during the game experience.

Strategy Analysis

For the strategy analysis, playtraces are compared to a standard algorithm for finding the MST of a graph called Prim's [32], which maintains and grows a tree by incrementally selecting the lowest cost edge encountered. Only edges that connect previously unconnected nodes are added to the MST solution and the process is repeated until all of the nodes are connected. If players solve the puzzles according to Prim's, it shows that *GrACE* helps to solve computational problems algorithmically. Other algorithmic approaches are possible, but Prim's was chosen as the golden standard algorithm because it directly aligns with our game mechanics.

We first performed a hierarchical cluster analysis to tease out different play styles used by *GrACE* players. Because the number of emergent strategies was not known *a priori* and hierarchical clustering operates without a predetermined number of clusters, each player run is hierarchically clustered based on the edit distances from one another and from the solution found by Prim's algorithm. Since sorting through qualitative data is time consuming, the idea is that clustering helps target the runs and level structures that need additional analysis with retrospective player sense-making.

For the retrospective player sense-making individual strategies were analyzed by retracing player steps by hand. This analysis involved redrawing on paper how players played each puzzle. Although this exercise seems tedious compared to watching a video of a player, the actual act of replaying the player steps on paper encourages us as researchers to imagine how players were making sense of the puzzle while playing, and therefore what strategies they were employing. In analyzing, the player steps were compared to a single run of Prim's algorithm on the same level.

Progression Analysis

Levels in *GrACE* are designed to increase in difficulty through the progression of the game, where difficulty is based on increasing numbers of nodes and edges. Levels with many failed player submissions indicate the challenge it posed to players. Progression visualizations can help evaluate whether the puzzles are progressively more difficult and where players first display some misunderstanding. We used a similar progression visualization method as Linehan et al. [25] who used it to detail how novel skills are introduced in a progressive manner.

An example of our player progression visualization method is shown in Figure 3. Each progression visualization represents a single playthrough of the game, but may not necessarily include completion of each level. For instance, Progression 1 in Figure 3 is an incomplete playthrough where the player never reached Level 6. Each dot represents a particular puzzle. The dot on the far left indicates Level 1 and on the far right Level 11. Lines connecting puzzles that arc upwards show forward progression in the game (e.g., moving from Level 1 to Level 2). Lines that arc downward are backward progressions (e.g., moving from Level 5 to Level 2) and indicate a replay of a previously completed puzzle. Loops starting and ending at the same dot are replays of the current level. The arc size (height and line width) and the boldness of the line are proportional to the number of times an answer for a particular level is submitted. For example, Progression 1 in Figure 3 has a large, wide loop on Level 5 illustrating that the level was attempted many times. Progression 2 in Figure 3 illustrates a player who progresses through the levels from Level 1 to Level 11 without replaying a current or previous level.

Participant	Pre	Post	Fun	Diff	Freq	Strategy
Red	NA	7	H	M	L	D
Orange	5	6	H	L	L	IT
Yellow	8	9	M	M	H	D
Green	8	8	L	H	M	IT
Blue	5	8	L	M	H	E
Indigo	7	5	H	L	M	D
Violet	3	4	H	H	M	D
Brown	8	9	M	L	L	D
Maroon	5	7	H	L	L	D

Table 1. The nine participants and their questionnaire results from the pilot study and predominant strategy from analyzing their gameplay. Questionnaire scores include pretest (Pre) and posttest (Post) scores (each ranges from 0 to 9), enjoyment (Fun) of *GrACE* (composite score ranging from 4 to 28), how difficult (Diff) they perceived *GrACE* (also from 4 to 28), how often they played games (Freq), and their predominant strategy (Strategy) during their initial playthrough of the game. Enjoyment and difficulty are listed in terms of Low < 14 (L), Medium = 14-20 (M), or High > 20 (H). Freq of game play ranges from 1 to 6 ("Never or Almost Never" to "Several Times a Day"). Freq is listed by Low = 1 or 2 (L), Medium = 3 (M) or High = 4 or 5 (H). There are three strategies: deliberate (D), exploratory (E), or iterative testing (IT).

Qualitative Analysis of Think Aloud Data

To better understand the player experience and address the emergent questions that arose from the aforementioned analyses, we analyzed the approximately hour-long recordings for each of the nine participants. Any audible utterances were transcribed and time-stamped, to later compare with game actions taken during the same time. The transcripts were investigated in two ways: 1) the talk was searched for evidence of strategy articulation during the trajectory of the players' experience and 2) the transcripts were used to align utterances to player action patterns at specific points of gameplay.

RESULTS

Results in this section explore the emergent strategies players exhibited and how well they can solve and learn MSTs through *GrACE*. Each player is represented by one of the following colors: Red, Orange, Yellow, Green, Blue, Indigo, Violet, Brown, Maroon. Each player's pretest, posttest, enjoyment

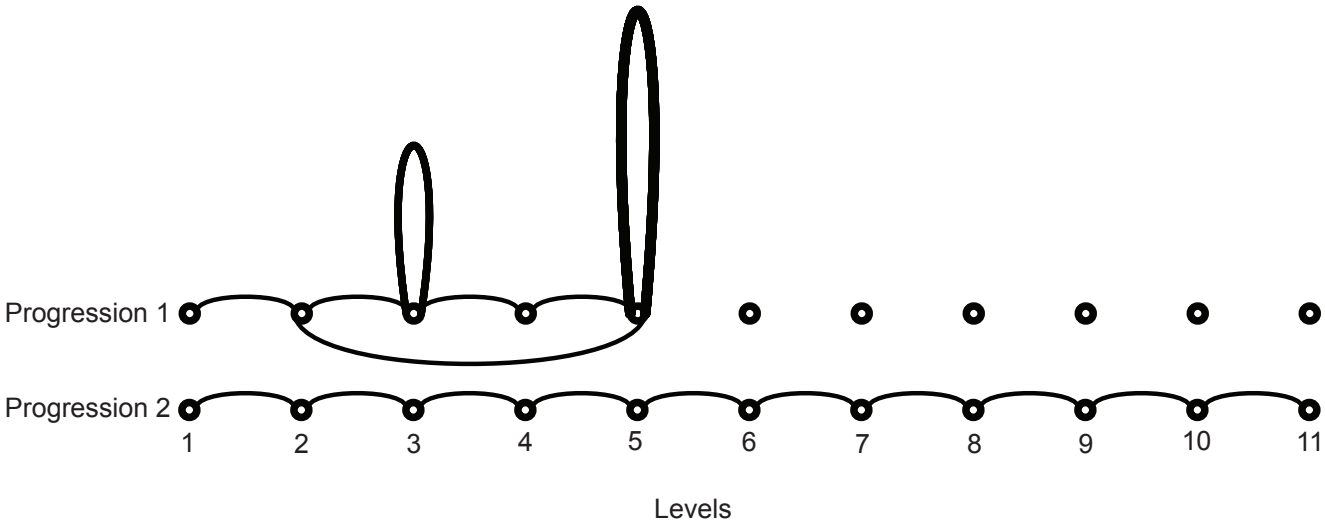


Figure 3. Example Player Progressions. Progression 1 shows a player progressing from level 1 to 2 to 3, repeating level 3 a few times, then progressing from 3 to 4 to 5, repeating level 5 many times, and finally going back to retry level 2. The player then quits playing at this point. Progression 2 shows a player progressing from level 1 all the way to level 11 without repeating any levels. Arcs above the dots indicate forward progression and arcs below the dots indicate a player going back to a previous puzzle.

of the game, assessment of its difficulty, frequency of game playing in general, and strategy is reported in Table 1. All but two individuals showed improvement from pretest to posttest (Green and Indigo). The table further shows diversity in how players experienced the game with no clear relationships between any of the measures, which indicates that these external measures may not provide the necessary insight to explain for these results, and that their actual play needs to be explored. The only interesting observation that can be inferred from the table is that it seems that players who play few games (Red, Orange, Brown, and Maroon) in their free time seem to find the game more fun and less difficult.

A broad overview of how players progressed through the game can be inferred from the progression visualizations in Figure 4. For instance, it is apparent that Green experienced difficulty grasping the game concepts as he completed only up to Level 5, and made multiple attempts from higher levels to go back and repeat lower levels. That idea is supported by the player's audio data expressing frustration and his expressed low enjoyment and high difficulty in playing the game. Similarly, though Blue never attempted to solve levels before her current level, her highest completed level is Level 9 and has a low enjoyment too. Blue, however, did seem to make the most visible improvement on the tests.

Evaluating successful players is less straightforward. For instance, because Red, Orange, Yellow, Indigo, Brown, and Maroon completed the game multiple times, they appear to have implemented an efficient strategy; however, further analysis of each player on each level leads to three emergent strategies called the 1) deliberate strategy, 2) exploratory strategy, and 3) iterative testing strategy. These strategies are indeed similar to previous research that grouped player strategies in response to challenge in the process of learning in single player games (trial and error, experiment, repetition, stop and think, and take the hint) [21]. A deliberate strategy typically indicates the player has a clear understanding of how to solve the puzzle by flagging as they move, and an iterative testing strategy typically indicates player experimentation or that the player is having difficulty solving the puzzles. An exploratory strategy is often an intermediate strategy where players explore large portions of the map before flagging any edges. The following analysis is broken up into analyzing these player behaviors at each level.

Play Analysis in Level 1

To help players understand basic gameplay in *GrACE*, Level 1 is based on the simple graph illustrated in Figure 6(a). It has two nodes, called Node 0 and Node 1 and one connection with a weight of one. Depending on their current position, players can either move left or right, and flag or unflag the path until their energy runs out. Solving this puzzle indicates a basic understanding of game mechanics.

All players eventually solve Level 1 with the least amount of moves, most on their first try. In fact, in the clustered representation of successful Level 1 attempts shown in Figure 5, each player has at least one run of Level 1 clustered with the computer's solution obtained through Prim's Algorithm. In this level, however, many also experiment with the mechanics. For instance, Yellow whose first run is solved near perfectly, begins experimenting with Level 1 on the second run. While in the first run, Yellow solves the puzzle by moving from Node 1 to Node 0, flagging the path and then submitting on the first try. In subsequent attempts, the player tests the game functionality figuring out whether it is necessary to move to each node and if a path can be flagged before moving to it.

Several users did have difficulty understanding these basic game mechanics, only reaching a solution after several attempts. Only three actions are necessary to solve this puzzle, yet Orange performed ten actions in the first two runs of the level, both unsuccessful. On the third try of Level 1, the cor-

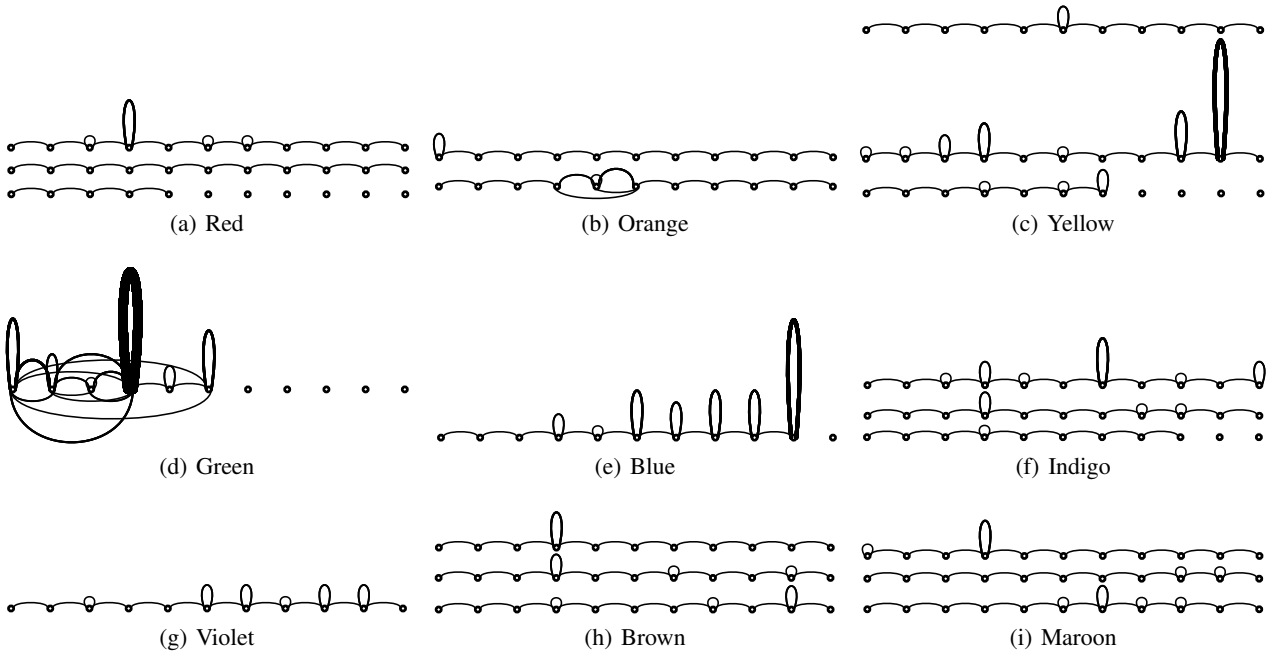


Figure 4. Progression visualizations over the 11 levels by player, where Level 1 is represented on the leftmost dot, and Level 11 the rightmost. Arcs above the dots indicate that players are advancing to the next level, while arcs starting and ending at same the dot represent that players are repeating that level. Then, arcs below the dots show players going back to solve levels that they have already solved. The arc size (height and line width) and the boldness of the line indicates the number of times a particular progression is repeated. In a set, the topmost progression is a player's first playthrough. The bottom progression is the final playthrough.

rect solution is found in nine tries, seemingly through trial and error. Eventually on the fourth run the puzzle is solved in three actions: move, flag, and submit.

Play Analysis in Level 3

Level 2 mostly reinforces the concepts that nodes must be visited, edges flagged and new information about a node's connections is discovered by visiting it. Level 3, on the other hand, challenges players to selectively choose edges to flag, important for identifying MSTs. Shown in Figure 6(b), a correct solution is moving between all the nodes, flagging connections between Nodes 0 and 1 and Nodes 1 and 2.

Though Red and Green start with a deliberate strategy, seemingly understanding flagging edges with minimum weights, once both encounter an "Out of Energy" error they retry the puzzle. During the second try in the same run, Red explores crossing the highest weight edge several times before running out of energy again, but eventually submits the correct answer. Similarly, Green submits the correct solution the second time Level 3 is played, but does not experiment with crossing the higher weight. Audio data suggests that Green grasps the goal of efficiency when exploring the graph. While playing Level 3, he utters "I'm not even sure how you do it...But, I don't need to go over there. Then, I run out of energy." Though he is likely still learning navigational affordances in the game, he appears to be attending to the energy constraints encountered by moving. Similarly, despite his actions Red acknowledges energy constraints, saying "...Too much time exploring."

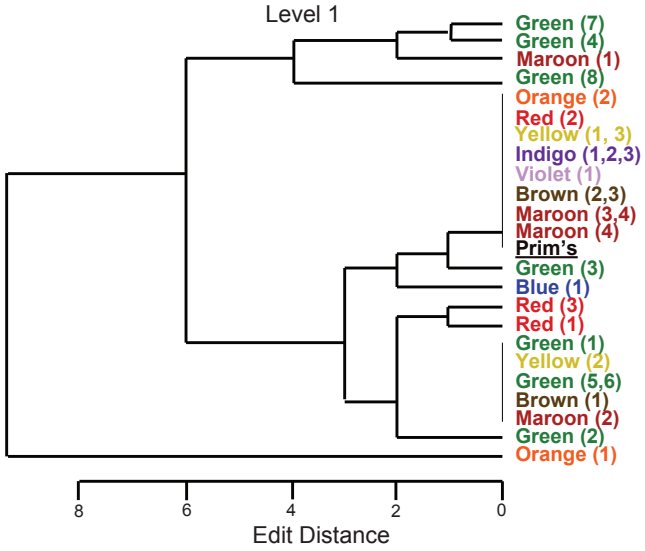
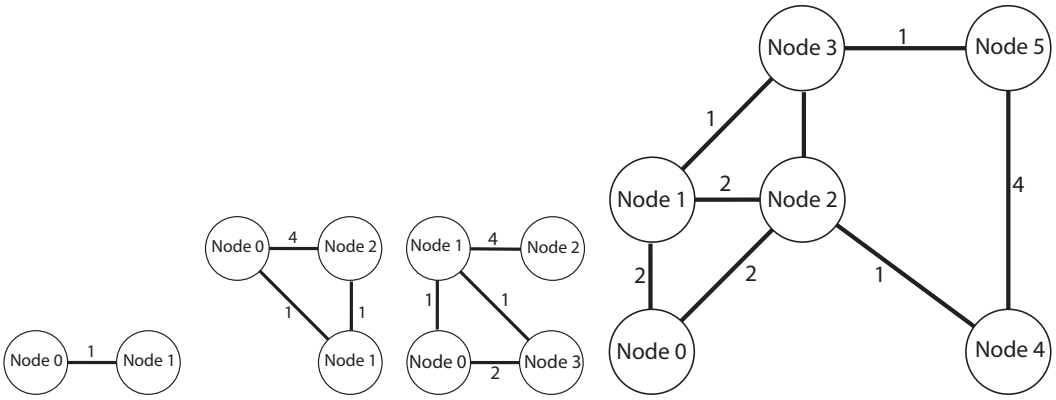


Figure 5. Clustering of *successful* playtraces for Level 1. Colored labels on the right indicate the player ID and run number. Some players completed one run (Blue and Violet) while others completed many (Green).



(a) Level 1 Graph (b) Level 3 Graph (c) Level 4 Graph (d) Level 6 Graph

Figure 6. Graph depictions of levels in *GrACE*. Levels in *GrACE* can be abstracted into graphs, where each node represents a burrow for the character to investigate for vegetables and each edge the path to a particular burrow. The goal is for players to plot a path so that each burrow is visited with the least amount of cost, or that is, the goal is to find a MST of the graph. The significant Levels 1, 3, 4, and 6 are depicted.

Yellow, Brown, Orange, and Indigo also all execute deliberate strategies while Violet and Maroon who had previously executed deliberate strategies start each run in Level 3 by exploring all of the nodes and then flagging them in an exploratory style. Blue previously implemented an iterate and test strategy, but for this level switched to exploratory.

This is the first level where players rerun levels to achieve the most efficient solution (i.e., one completed in the least amount of steps). Only Yellow and Green find the solution on their first attempts. Most players eventually converge on the Prim's solution, yet Orange and Blue never find a solution comparable to Prim's.

Play Analysis in Level 4

With a total of four nodes and four edges compared to Level 3's three nodes and three edges, Level 4 shown in Figure 6(c) produces even greater challenges to the players. Furthermore, in this level it is possible to select two edges not connected to the same node. It is evident from the progression visualizations that Level 4 presents the first big hurdle for players with most having an increased number of attempts on their first playthrough of the game. The correct solution involves flagging a path of weight 4 between Nodes 1 and 2, and choosing not to connect Nodes 0 and 3 with weight 2. The other two paths with weight 1 should be flagged instead.

Initially, players struggle to only connect the lowest weight edges. For example, Blue and Violet begin by flagging all of the edges, while Maroon connects all of the nodes and excludes an edge, but does so without regard for which has the lowest weight. Interestingly, on a second playthrough of the game, Maroon solves the puzzle in a number of steps similar to that calculated by Prim's.

Perhaps because all lowest cost weights in previous levels were 1, it is possible that some players faced difficulty flagging the only connection to Node 2, which has a weight of 4. Everyone eventually solves Level 4, however, only Yellow, Brown, and Maroon present solutions comparable to Prim's. Of these players, only Yellow solves it like Prim's on the first try. These players perform the best throughout the remaining levels, with Yellow and Brown scoring perfectly on the posttest, and Maroon improving more than others (except for Blue). Interestingly, Maroon found it more enjoyable than Yellow and Brown, possibly because she learned more.

Play Analysis in Level 6

Pictured in Figure 6(d), Level 6 is a graph with six nodes and eight edges, and is the first to present a node connected to four other nodes, testing the player's ability to choose which edge for the node should be flagged. Level 5, the last that all players were able to solve, presented at most three edge selections at any given node and had two fewer nodes overall.

Even Yellow who had previously solved each level easily and optimally took three tries to solve the puzzle. Starting at Node 2, on her first try, Yellow moved along a shortest path to Node 4, flagged it, and then moved to Node 5 across a path of weight 4. She flagged the higher weight path after visiting Node 5, meaning this weight was chosen after seeing the cheaper connection to Node 5 from Node 3. From Node 5, she went to Node 3, flagged the path and then moved to Node 1. After moving from Node 1 to Node 0, Yellow ran out of energy and restarted the level.

Since Yellow previously seemed to understand the necessity of choosing the lowest cost path, it is notable that she chose the weight path of value 4 connecting Nodes 4 and 5. Though she eventually finds the correct answer, the confusion is reflected across other players. For instance, although Violet starts deliberately moving and flagging, after the first submission fails, he explores and iterates and tests until six submissions later he finds the right answer.

When comparing to the answer found by Prim's algorithm, for this level the edit distance from any given solution is greater than zero, with Brown's second try and Maroon's second and third tries being the closest. However, Orange, Yellow, Indigo, and Brown were all eventually able to complete Level 6 with only one more action than Prim's algorithm.

Play Analysis Levels 7 through 11

Levels 7 through 11 continue to reinforce the concepts introduced by previous levels, challenging the players' existing strategies on more difficult puzzles. Players who have already discovered the deliberate strategy described in Table 1 have little trouble completing these levels, while those with exploratory and iterative testing strategies struggle with the additional nodes and edges.

Level 9 is the last level Blue completes, perhaps because Level 10 jumps to eight nodes, 13 edges, and a maximum connectivity of six, versus the six nodes, ten edges, and maximum connectivity of five seen in Level 9. The average connectivity of a node decreases slightly (3.30 to 3.25), but the highest connectivity of a node increases from 5 to 6.

Inspecting the clusters indicates that Red, Orange, and Violet begin to diverge from Prim's algorithm at Level 7, and Red and Orange continue this pattern until the end. Interestingly, Red and Orange sat next to each other while playing the game and were frequently talking to each other during game play, although they were instructed to solve the problems on their own—a fact highlighted only through audio data. Importantly, their similarity in play was detectable in their game actions. Their conversation included critique of the game—first Red then Orange posited that they found "a bug." Red explained that the game "says I'm wrong when it in actuality I was right." Later, he repeated "the last few levels I had the right answer and it said try again." Occasionally Orange asks Red for help and Red is continuously vocal in discussing his critiques with researchers. Through clustering of their playtraces and retrospective player sense-making, the difficulty they faced solving levels in *GrACE* is evident.

Indigo, Maroon, Brown, and Yellow continued to solve the puzzles in a manner similar to Prim's algorithm. Interestingly, Violet seemed to be implementing a deliberate strategy followed by an amount of guessing and checking, and by Level 11 the strategy was clearly cemented. Violet's progression in Level 10, began to hint that the main ideas were being learned.

In fact, our analysis suggests that over the course of the game, most students learn to visit all of the nodes in the MST, prefer traveling on shorter paths to avoid running out of energy, the concept of a spanning tree, and how to find the MST.

DISCUSSION

Our results highlight that including game data in analysis can shed light onto findings from traditional measures in addition to giving a better understanding of what players actually do. However, in some cases it raises more questions such as in the interesting case of Blue. In this section, we focus on discussing our findings for what made levels more or less difficult and the emergent player strategies and their implications for educational game design.

Although this work represented analysis of data for a particular game, future work will focus on its scalability to larger data sets and different games. For instance, with data from more players, an archetype clustering [5] can be performed helping sort the data into small enough subsets to perform meaningful hierarchical clustering.

Level Complexity and Progression

To avoid boredom and frustration, and to support learning effectively, it is important that the difficulty of levels increases steadily over time. However, through the progression analysis in this paper it is apparent that Level 6 was the first to present such a significant challenge to players that some could not successfully complete it. By tracing through individual player actions and examining the level structure (i.e., its underlying graph), part of the challenge appeared to be the increase in the number of nodes and edges from that in Level 5 (i.e., from four nodes to six nodes and five edges to seven).

Interestingly, while Level 11 also increased the number of nodes and edges from that in Level 10, this increase posed less of a challenge to players. While an explanation for this difference could in part be that those who were struggling most never reached Levels 10 and 11, further analysis across all levels in the progression reveals that in fact the biggest predictor of challenge lies in the number of choices available to players at a given node (i.e., the number of edges connected to a node). The number of edges connected to the node with the greatest amount of choices represents the maximum connectivity of the underlying graph. Although increasing the number of nodes and edges can lead to increased maximum connectivity, alone it is only a secondary indicator of difficulty. In fact, Levels 4, 6, 8, and 10 all saw increased maximum connectivity and were the levels players struggled with the most. Before our analysis, it was thought level difficulty could be sufficiently represented by the number of nodes and edges; however, a detailed trace analysis revealed the primary indicator of level difficulty as the graph's maximum connectivity.

Although the issue of maximum connectivity of a graph is specific to graph-based games, the process of discovering this phenomenon is not. We suggest that designers of educational games add into their development cycle a phase exploring where players get stuck, identifying the factors that contribute to challenges for those players, and then modifying and retesting the level progression. Through application of this procedure, designers can adjust the progression in an effort to reduce player frustration and increase learning.

That said, it is an open question to designers if reducing frustration needs to be aimed for. Blue is an interesting case. She did not enjoy the game and did not progress as far in the game as others, yet made the most visible improvement. Although Blue is a single case, in our other analyses with *GrACE* [9] we found a similar paradox: that players who did not enjoy it seemed to be the ones most benefiting from it. Regardless of how to deal with frustration, it will be beneficial to designers to understand what makes their game really difficult.

Player Strategy

Player strategies were determined through the mixed-methods approach described in this paper. We discuss how we determined such strategies and what they mean in the context of other findings. In addition, we discuss the implications for educational game design, and focus on the importance of game mechanics in encouraging (or hindering) learning and determining the appropriate measures of success.

Determining Strategies

Whether a player solved a level is important information in analyzing an educational game, but to explore learning it is crucial to understand *how* players solved the problem. Although player strategies are often discerned through laborious qualitative studies [21], the mixed-methods approach presented in this paper reduced time spent analyzing the qualitative data by allowing us to focus on critical game events and problematic levels. Clustering players' game actions combined with playtrace and audio analysis led us to determine that six of the nine players employed a deliberate strategy while two used an iterative testing method and one used an exploratory strategy.

When characterizing player strategies, it is difficult to conclusively state why particular actions were taken, but some evidence exists as to why player experiences diverged. For example, there appear to be trends related to player engagement as shown in Table 1. Player strategy appears independent to either their enjoyment or perception of difficulty. Looking at Orange and Green, who both took iterative approaches, Orange had a low rating of difficulty with a high rating of enjoyment. Green had a high rating of difficulty and a low rating of enjoyment. At least for the iterative approach, strategy development does not appear related to enjoyment or difficulty; however, both players had a gain of either 0 or 1, and so it may be the case that iterative strategies did not support learning. These are potential trends that need to be explored further.

Although surprising, improvement in score does not appear to be driven by enjoyment of the game. Counter-intuitively, those who played games more often rated the game at a medium or high difficulty, and medium and low enjoyment. One might expect that frequent gamers might enjoy a new game, and would find it less difficult than other players. These patterns require further investigation, but one hypothesis is that the players may have perceived the game as an educational task, rather than a true game, and the very logical nature of the task compared to the games they usually play, may have affected their engagement, learning (e.g., despite low engagement, Blue was one of two participants with high frequency of gaming, but was also the participant who gained most from pre to post), and potentially strategy development.

Interleaving Game Mechanics, Strategy, and Learning

Another important finding is the impact of game mechanics on learning. For example, Violet initially began with an efficient strategy, but was often confused by the game mechanic that required visiting all burrows before submitting a correct answer. Even though he sometimes began with a deliberate and correct strategy, when he failed to visit all burrows, his answer was judged as incorrect leading him to doubt his original strategy. Educational games that allow multiple strategies to lead to success should carefully consider their game mechanics' interactions with player strategies and how those impact learning. It is important to support all valid strategies and give proper feedback at failure points, which may vary based on strategy.

Metrics for Success

Another issue is developing appropriate metrics for success. As seen in previous playtrace research [2], trace length (number of actions) was insufficient to determine difficulty (or player success) of levels. In *GrACE*, some players successfully completed levels in as few actions as possible whereas others adopted a more exploratory strategy and would successfully complete levels using more actions than required. Neither of these strategies coincided with a more successful approach as they both successfully completed the levels.

Additionally, players could be equally efficient but have playtraces clustered separately. This was seen on Level 6 between Orange, Yellow, Indigo, and Brown. Only two were grouped with Prim's algorithm even though they all had identical efficiency and trace length. Although our original intent was to allow players to find their own algorithm since many exist for solving the MST problem, it became evident upon playtrace analysis that an additional measure of success (e.g., number of actions taken) is required during gameplay to encourage a more algorithmic approach than guess-and-check. In our game, the measure of success for efficiency—Scout's energy—was not reinforced enough since we didn't want to limit players to only one algorithm. Because of this, players would hardly ever run out of energy even with an iterative testing strategy. We encourage other algorithmic thinking game designers to implement similar measures of success to ensure a player must be algorithmic to be successful. In fact, this approach can be extended to any game where designers want players to exhibit a particular behavior.

CONCLUSION

In this research we have detailed the analysis from nine individuals playing *GrACE*, an educational puzzle game emphasizing algorithmic thinking. Our research outlines insights gained through the use of a mixed-methods strategy for gameplay analysis that triangulates data from four perspectives: think-aloud voice recordings, qualitative playtrace analysis we call *retrospective player sense-making*, level progression visualizations and quantitative clustering on player actions. We have shown how a mixed-methods approach can be applied to an educational game and how it helped us gain insights that may have been overlooked or incorrectly inferred without game analytics or using a single method to analyze playtraces. Finally, data-driven design and evaluation of an educational game has allowed us to detect problem areas early in the design phase where steps can still be taken to correct shortcomings. We hope that future game designers—educational and otherwise—take into consideration the benefits of mixed-methods approaches to game analytics and develop appropriate metrics for success that will ensure the game designer's goals are met.

ACKNOWLEDGMENTS

We want to thank the Northeastern STEM Center for Education for integrating our pilot study in their summer program and Christopher Clark and Oskar Strom who helped create the game. This material is based upon work supported by the National Science Foundation under Grant No. 1422750.

REFERENCES

1. Mike Ambinder. 2009. Valve's approach to playtesting: The application of empiricism.
2. Erik Andersen, Sumit Gulwani, and Zoran Popović. 2013. A Trace-Based Framework for Analyzing and

Synthesizing Educational Progressions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 773–782.

3. Erik Andersen, Yun-En Liu, Ethan Apter, Francois Boucher-Genesse, and Zoran Popović. 2010. Gameplay Analysis Through State Projection. ACM, 1–8.
4. CodeCombat. 2016. CodeCombat: Learn to Code by Playing a Game. (2016).
5. Adele Cutler and Leo Breiman. 1994. Archetypal analysis. *Technometrics* 36, 4 (1994).
6. Phillip DeRosa. 2007. Tracking player feedback to improve game design. *Gamasutra* (2007).
7. Nicholas Duchenaut, Nick Yee, Eric Nickel, and Robert J. Moore. 2006. Building an MMO with mass appeal: A look at gameplay in World of Warcraft. *Game Developer Magazine* 1 (2006), 281–317.
8. Magy Seif El-Nasr, Anders Drachen, and Alessandro Canossa. 2013. *Game analytics: Maximizing the value of player data*. Springer Science & Business Media.
9. Yetunde Folajimi, Britton Horn, Jacqueline Barnes, Amy Hoover, Gillian Smith, and Casper Harteveld. 2016. A Cross-Cultural Evaluation of a Computer Science Teaching Game. In *Proceedings of Games+Learning+Society*. ETC Press, Pittsburgh, PA.
10. Tracy Fullerton. 2008. *Game Design Workshop: A playcentric Approach to Creating Innovative Games* (2nd ed.). Morgan Kaufmann.
11. Bradley S Greenberg, John Sherry, Kenneth Lachlan, Kristen Lucas, and Amanda Holmstrom. 2010. Orientations to video games among gender and age groups. *Simulation & Gaming* 41 (2010), 238–259.
12. Erik Harpstead and Vincent Aleven. Using Empirical Learning Curve Analysis to Inform Design in an Educational Game. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY 2015)*. ACM Press, 197–207.
13. Erik Harpstead, Christopher J MacLellan, Vincent Aleven, and Brad A Myers. 2015. Replay Analysis in Open-Ended Educational Games. In *Serious Games Analytics*, Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler (Eds.). Springer International Publishing, 381âĂŞ399.
14. Casper Harteveld. 2011. *Triadic game design: Balancing reality, meaning and play*. Springer Science & Business Media.
15. Casper Harteveld. 2012. *Making Sense of Virtual Risks*. IOS Press.
16. Casper Harteveld, Gillian Smith, Gail Carmichael, Elisabeth Gee, and Carolee Stewart-Gardiner. 2014. A Design-Focused Analysis of Games Teaching Computer Science. In *Proceedings of Games+Learning+Society 10*. Madison, WI.
17. Casper Harteveld and Steven Sutherland. 2015. The goal of scoring: Exploring the role of game performance.. In *Proceedings of the 2015 ACM Conference on Computer-Human Interaction*.
18. Eric Hazan. 2013. Contextualizing Data. In *Game Analytics*. Springer, 477–496.
19. David Hilbert and David Redmiles. 2000. Extracting usability information from user interface events. *Comput. Surveys* 32, 4 (2000), 384–421.
20. Britton Horn, Christopher Clark, Oskar Strom, Hilery Chao, Amy J. Stahl, Casper Harteveld, and Gillian Smith. 2016. Design insights into the creation and evaluation of a computer science educational game. In *Proceedings of the 47th ACM Technical Symposium on Computer Science Education (SIGCSE)*. ACM Press, Memphis, TN.
21. Ioanna Iacovides, Anna L Cox, Ara Avakian, and Thomas Knoll. Player Strategies: Achieving Breakthroughs and Progressing in Single-player and Cooperative Games. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play (CHI PLAY 2014)*. ACM Press, 131–140.
22. Amanda Lenhart and Pew Internet & American Life Project. 2008. *Teens, Video Gaming and Civics*. Technical Report.
23. Vladimir I Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics - Doklady* 10, 8 (1966), 707–710.
24. Frederick WB Li and Christopher Watson. 2011. Game-based concept visualization for learning programming. In *Proceedings of the third international ACM workshop on Multimedia technologies for distance learning*. ACM, 37–42.
25. Conor Linehan, George Bellord, Ben Kirman, Zachary H Morford, , and Bryan Roche. Learning Curves: Analysing Pace and Challenge in Four Successful Puzzle Games. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play (CHI PLAY 2014)*. ACM Press, 181–190.
26. Yun-En Liu, Erik Andersen, Richard Snider, Seth Cooper, and Zoran Popović. 2011. Feature-based Projections for Effective Playtrace Analysis. In *Proceedings of the 6th International Conference on Foundations of Digital Games (FDG '11)*. ACM, New York, NY, USA, 69–76.
27. Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler. 2015. *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*. Springer.
28. Truong-Huy Dinh Nguyen, Magy Seif El-Nasr, and Alessandro Canossa. 2015. Glyph: Visualization Tool for Understanding Problem Solving Strategies in Puzzle Games. *Foundations of Digital Games (FDG)* (2015).
29. Jackie O'Kelly and J Paul Gibson. 2006. RoboCode & problem-based learning: a non-prescriptive approach to teaching programming. *ACM SIGCSE Bulletin* 38, 3 (2006), 217–221.

30. Joseph C Osborn, Ben Samuel, Joshua Allen McCoy, and Michael Mateas. 2014. Evaluating play trace (dis)similarity metrics. In *10th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

31. Mikki H Phan, Jo R Jardina, Sloane Hoyle, and Barbara S Chaparro. 2012. Examining the Role of Gender in Video Game Usage, Preference, and Behavior. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications, 1496–1500.

32. Robert C. Prim. 1957. Shortest Connection Networks And Some Generalizations. *Bell System Technical Journal* 36 (Nov. 1957), 1389–1401.

33. Ramon Romero. 2008. Successful instrumentation: Tracking attitudes and behviors to improve games. In *Game DeveloperâĂŹs Conference*.

34. Angel Serrano-Laguna, Javier Torrente, Pablo Moreno-Ger, and Baltasar FernÃąndez-Man. 2014. Application of Learning Analytics in Educational Videogames. *Entertainment Computing* 5 (2014), 313–322.

35. David Williamson Shaffer and James Paul Gee. 2012. The Right Kind of GATE: Computer games and the future of assessment. *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research. Charlotte, NC: Information Age Publishing* (2012).

36. Valerie Shute and Matthew Ventura. 2013. *Stealth assessment: Measuring and supporting learning in video games*. MIT Press, Cambridge, MA.

37. George Siemens and Ryan S J d Baker. 2012. Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. ACM, 252–254.

38. Catherine A Sugar and Gareth M James. 2003. Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach. *J. Amer. Statist. Assoc.* (2003), 750–763.

39. Robert L Thorndike. 1953. Who belongs in the family? *Psychometrika* 18 (Dec. 1953), 267–276.

40. Allen Tucker, Fadi Deek, Jill Jones, Dennis McCowan, Chris Stephenson, and Anita Verno. 2006. *ACM K-12 CS Model Curriculum*. Computer Science Teachers Association.